
A Machine Learning Predictive Model for Determining Credit Risks In Ethiopian Microfinance Institutions

Abebaw Bogale Ejigu^{1*} Dr. Solomon Demissie ²

**^{1*} Collage of Natural and Computational Science, Department of Computer Science,
Kotebe University of Education**

**² Collage of Natural and Computational Science, Department of Computer Science,
Kotebe University of Education**

Email: ² soldave99@yahoo.com

Corresponding Email: ^{1*} abebawbg47@gmail.com

Abstract

The rapid growth of financial data due to globalization and technological advancement has created a need for automated techniques in financial decision-making. Traditionally, loan evaluations in microfinance institutions relied heavily on the subjective judgment of loan officers. Previous studies on credit risk analysis used limited datasets and features, resulting in lower accuracy. This research aims to enhance credit risk prediction by building and evaluating machine learning models using a more extensive dataset from various microfinance institutions. The study involved preprocessing a dataset of 13,148 records and testing four algorithms: LightGBM, SVM, Decision Tree, and KNN. The results showed that the LightGBM model achieved the highest accuracy at 99.97%, followed by SVM at 97.30%, Decision Tree at 96.79%, and KNN at 94.46%. The study concludes that adopting machine learning models, particularly LightGBM, and implementing standardized data warehouses can significantly minimize credit risks in microfinance institutions. Future research should explore larger datasets, more attributes, and new algorithms to further improve outcomes.

Keyword: Credit Risk Prediction, Machine Learning, Microfinance Institutions, LightGBM, Financial Data Analysis

1. Introduction

Ethiopia's credit system and market consist of various organizations, including banks, non-bank credit institutions, credit cooperatives, and Microfinance Institutions (MFIs). These entities play a crucial role in providing financial services to individuals and businesses that do not meet traditional bank requirements. MFIs, in particular, offer small loans to people with limited access to formal banking services. This is vital for improving living standards, employment, and entrepreneurship in Ethiopia. The country's development strategy emphasizes the establishment of sustainable MFIs to serve a large portion of the poor population. Although informal financing and NGO credit schemes have existed in Ethiopia for years, the legal framework for MFIs was established with Proclamation 40/1996. [1]

Credit quality has become increasingly important in MFIs' lending decisions, with credit scoring techniques used to assess borrowers' creditworthiness. However, the lack of a centralized credit history database poses challenges for accurately determining an individual's creditworthiness, especially in micro-lending environments like those in Ethiopia. Traditional methods of credit risk assessment, such as linear or logistic regression, have limitations, and machine learning techniques have emerged as powerful tools in various fields, including finance. [2] By recognizing patterns in financial data, machine learning models can improve credit risk assessment, enabling MFIs to make more informed lending decisions. This research focuses on developing machine learning models to predict potential borrowers' creditworthiness in Ethiopian MFIs, contributing to more effective credit risk management.

Statement of Problems

Microfinance Institutions (MFIs) in Ethiopia face challenges in managing credit risk due to the absence of a centralized credit history database and limited information on borrowers' creditworthiness. Traditional methods of evaluating borrowers, often based on the judgment of loan officers, are insufficient for accurately assessing lending risks. The lack of comprehensive data and reliance on subjective assessments increase the likelihood of loan defaults, which can jeopardize the financial stability of MFIs. This research aims to address these challenges by developing machine learning models that can predict borrowers' creditworthiness, thereby enhancing the decision-making process in Ethiopian MFIs and reducing the risk of default.

Objective of the Study

The primary objective of this study is to develop a machine learning model that predicts potential borrowers for Microfinance Institutions (MFIs). Specific objectives include reviewing relevant literature, preparing datasets with key attributes, identifying critical features for credit risk prediction, building and evaluating machine learning models, conducting experiments using specific tools, and providing recommendations for future research.

Literature Review

Salame found decision trees to be the most suitable model for predicting loan default risk in an agricultural financial institution. [3] Sirikulvadhana demonstrated that clustering techniques in data mining can enhance financial auditing by uncovering patterns missed by traditional audit software. [4] Jia et al. identified decision trees as the best technique for building a loan risk assessment system for subprime lenders. [5] Koyuncugil and Ozgulba highlighted the accuracy of decision trees for risk assessment in SMEs. [6] Raju, Bai, and Chaitanya emphasized the importance of historical data in risk assessment, using decision trees and neural networks for client retention in banking. [7] Denekeew A. improved CRM at Ethiopian Airlines using K-means clustering and J48 for customer segmentation. [8] Sara Worku created a credit analysis model at Addis Credit and Savings Institution with decision trees, Naive Bayes, and PART, but was limited by a small dataset. [9] Khandani et al. showed that machine learning models could reduce customer credit risk losses by 6% to 25%. [10] Yap et al. found no significant performance difference between logistic regression, decision trees, and credit scorecards in detecting subscription defaulters. [11] Zhao et al. achieved 87% accuracy using MLP neural networks for credit score estimation and noted that increasing hidden units could improve accuracy. [12]

Gap Analysis

The gap analysis reveals that previous research on credit risk prediction in microfinance institutions has primarily focused on organizational, business, and customer characteristics using limited datasets and common algorithms like Decision Trees and Naive Bayes. These studies have not adequately addressed borrower-specific attributes such as collateral, credit history, and educational status. This research aims to fill this gap by developing a classification model that incorporates these novel

features, which have not been explored in previous studies. By utilizing less commonly used algorithms, such as LightGBM, alongside traditional methods, the study seeks to improve the accuracy and relevance of credit risk prediction for Ethiopian Microfinance Institutions. This approach is data-driven and targets real-world credit risk factors to enhance decision-making and reduce default rates.

2. Research Design

This research addresses the problem of identifying the appropriate predictor variables in classification. The purpose of the classification is to predict if the borrower is likely to active or default on their loan based on the information gathered throughout the loan application process. The design includes data collection, data preprocessing, and data analysis; feature selection, splitting data in to test and training model creation using SVM, KNN, decision tree and LightGBM algorithms; and data visualization using selected algorithms. Based on this the CRISP-DM technique has been used to design this experiment.

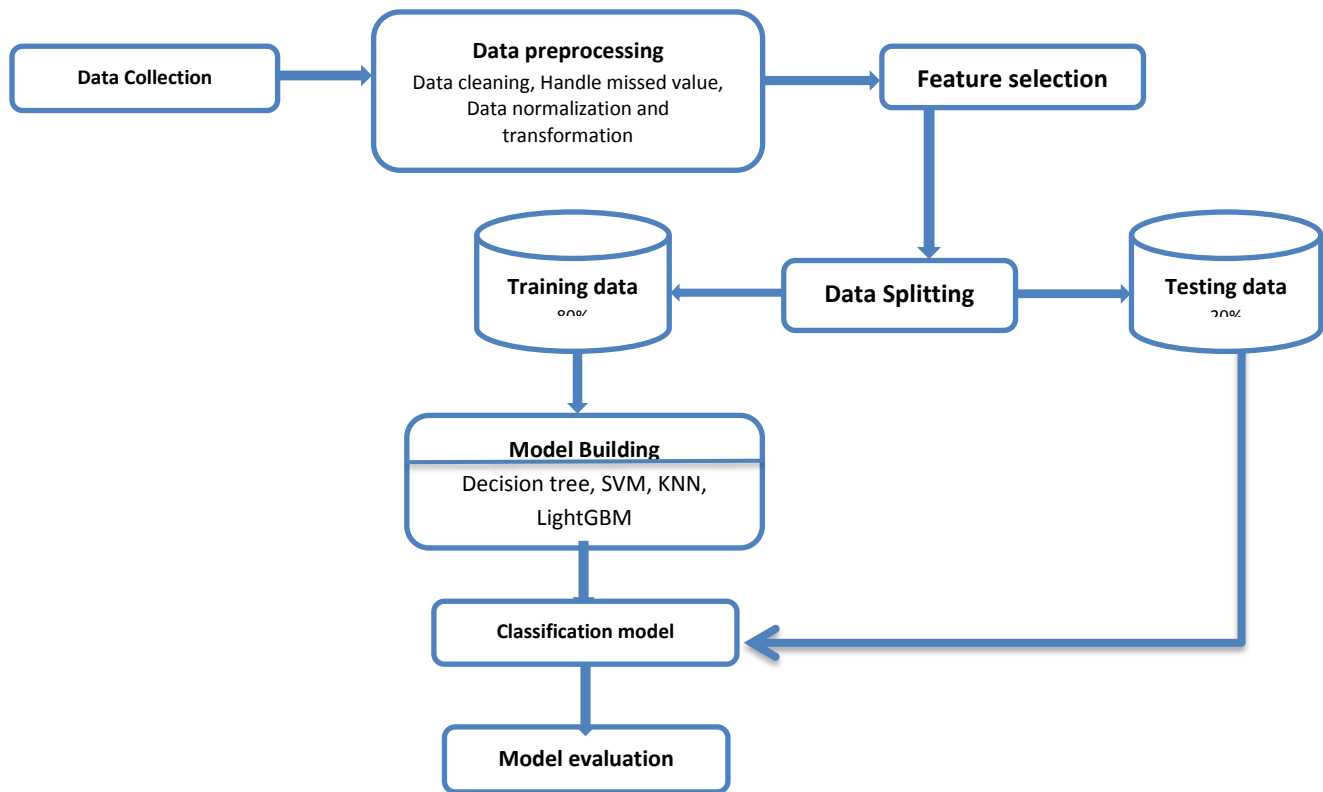


Figure 6: Research design

Data Collection

Data collection involved in-depth interviews with domain experts to determine attribute feature selection and understand business processes in Ethiopian Micro finance Institutions, which is facing high default rates and failing to meet the international standard of a 5% portfolio at risk. [13] The study collected data spanning from 2016 to 2021, focusing on borrower characteristics, credit attributes, and business features. The dataset includes 13,148 records with 14 attributes, including a dependent variable indicating default status. This comprehensive data collection, guided by expert advice, aims to address the institution's credit risk issues and improve prediction accuracy.

Data Cleaning

Data cleaning involves removing or correcting erroneous, incomplete, irrelevant, or redundant data to enhance analysis accuracy. This process includes addressing issues like spelling errors, standardizing data, and handling missing values. For attributes such as age, sex, and marital status, records with missing values were removed, leaving 13,148 entries for further analysis.

Data Preprocessing

Data preprocessing prepares data for machine learning by addressing issues like noisy data, missing values, and deriving new fields. This step involves cleaning, summarizing, and refining the data, ensuring it is suitable for analysis. The data is saved in a .CSV file in Microsoft Excel, optimizing it for accurate and efficient classification.

No	Attribute Name	Data type	Description
1	Loan ID	Numeric	the Loan ID No
2	Customer Name	Character	Name of customers
3	Sub city	Character	Name of the sub city
4	Wereda	Character	Name of woreda
5	Date	Numeric	Date of credit application
6	Security saving	Numeric	Mandatory Saving improve credit repayment
7	Voluntary saving	Numeric	Saving voluntarily in addition to repayment
8	Credit History	Numeric	Number of times the customer get credit
9	Saving balance	Numeric	Saving amount deposited
10	Loan amount	Numeric	Amount of Loan borrowed
11	Paid principal	Numeric	Paid balance
12	Outstanding	Numeric	Remaining unpaid balance

13	Loan Status	Numeric	Profitability level of customers
14	Age	Numeric	Age of customers
15	Sex	Character	Sex of customers
16	Marital Status	Character	Marital Status of customers
17	Educational Level	Character	Education status of borrower
18	Other source of income	Numeric	Extra source of money
19	Collateral	Numeric	Security for repayment of a credit
20	Business Sector	Character	Business type
21	Loan Term	Numeric	Loan duration in month

Table 1: Provides a description of the dataset's features

Data Transformation

Data transformation adjusts and standardizes data to make patterns more comprehensible before analysis. In this study, ages were filtered to 18 and above, educational status was categorized and numerically coded, and loan status and other attributes were similarly encoded for consistency. Attributes like sex, marital status, credit history, and income sources were transformed into numerical values for clearer analysis and to facilitate pattern recognition.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Saving_Balance	Loan_Amount	Paid_principal	Outstanding	Loan_Status	Age	Sex	Marital_Status	source_of_i	Educational_Level	Colateral	Business_Sector	Loan_Term	Credit_History
2	1000	5000	0	5000	0	30	1	1	0	1	197	3	24	2
3	1000	5000	5000	0	1	38	0	2	1	3	197	3	24	2
4	1000	5000	5000	0	1	37	1	2	1	3	197	3	24	1
5	1000	5000	5000	0	1	27	1	1	1	2	93	0	18	1
6	1000	5000	5000	0	1	42	0	2	1	3	93	0	24	1
7	1350	15000	0	15000	0	35	1	2	0	3	93	0	24	1
8	1000	5000	5000	0	1	27	1	1	1	2	93	0	24	1
9	1000	5000	3500	1500	0	29	1	1	0	1	93	0	12	1
10	600	3000	3000	0	1	42	0	2	1	3	93	0	12	1
11	6000	30000	30000	0	1	28	1	1	1	1	93	0	24	1

Figure 1: Transformed dataset

Attribute Selection

Attribute selection aims to identify the most relevant features for building effective machine learning models. Techniques used include Fisher's Score for ranking variables and correlation coefficients to assess the linear relationship between variables. The goal is to choose attributes that are highly correlated with the target variable while avoiding redundancy. For the study, attributes related to borrowers' credit risk were selected based on their relevance, and less relevant data such as addresses were excluded to streamline the dataset for analysis. [14]

Correlation Analysis

Correlation analysis measures the strength of relationships between variables. A perfect correlation means one variable predicts another accurately. [15] The study found that several attributes like saving balance, loan amount, and credit history have strong correlations with each other, indicating their importance in predicting credit risk.

Data Set Description

The dataset consists of 13,148 records from different Ethiopian Micro finance Institutions, each with 14 attributes. These attributes include borrower characteristics and credit details, used to analyze and predict credit delinquency.

No	Attribute Name	Transformed data type
1	Saving_Balance	Numeric
2	Loan_Amount	Numeric
3	Paid_principal	Numeric
4	Loan_Status	Numeric
5	Sex	Numeric
6	Marital_Status	Numeric
7	Other_source_of_income	Numeric
8	Educational_Level	Numeric
9	Colateral	Numeric
10	Business_Sector	Character
11	Loan_Term	Numeric
12	Credit_History	Numeric

Table 2: Selected attributes of the dataset

Data Quality Assurance Methodology

Data quality assurance involves identifying and addressing abnormalities through data profiling, removing outdated information, and cleansing the data to ensure accuracy and completeness. This process is crucial for maintaining the integrity and value of data throughout its lifecycle. For the research, the dataset from different Micro finance Institution was validated for completeness, accuracy, and relevance before analysis to ensure it contained all necessary attributes for effective credit risk prediction. [16]

Predictive Models

This study employs four prediction techniques: K-nearest neighbor (KNN), Support Vector Machines (SVM), Decision Tree, and LightGBM. Each model is developed with mutually exclusive branches and uses a hyperplane for predictions. Hyper parameters are optimized using systematic, random searches, or heuristics. [17] The data is split into 80% training and 20% testing sets, with stratification to ensure representative distribution. The models are evaluated with 5-fold cross-validation and performance is assessed both with and without hyper parameter tuning.

Proposed Model

When the algorithms were chosen, the model was run using the parameters of the algorithms. Using classification and prediction methods, several tests with various parameters would be carried out. In Ethiopian Micro Finance Institutions dataset, a predictive model for predicting individual customer behavior would be used. While building models, the approach is expected to learn specific patterns in the dataset, which are then applied to new datasets. The model is expected to forecast whether or not a credit applicant would default on a loan. The following diagram depicts the system architecture

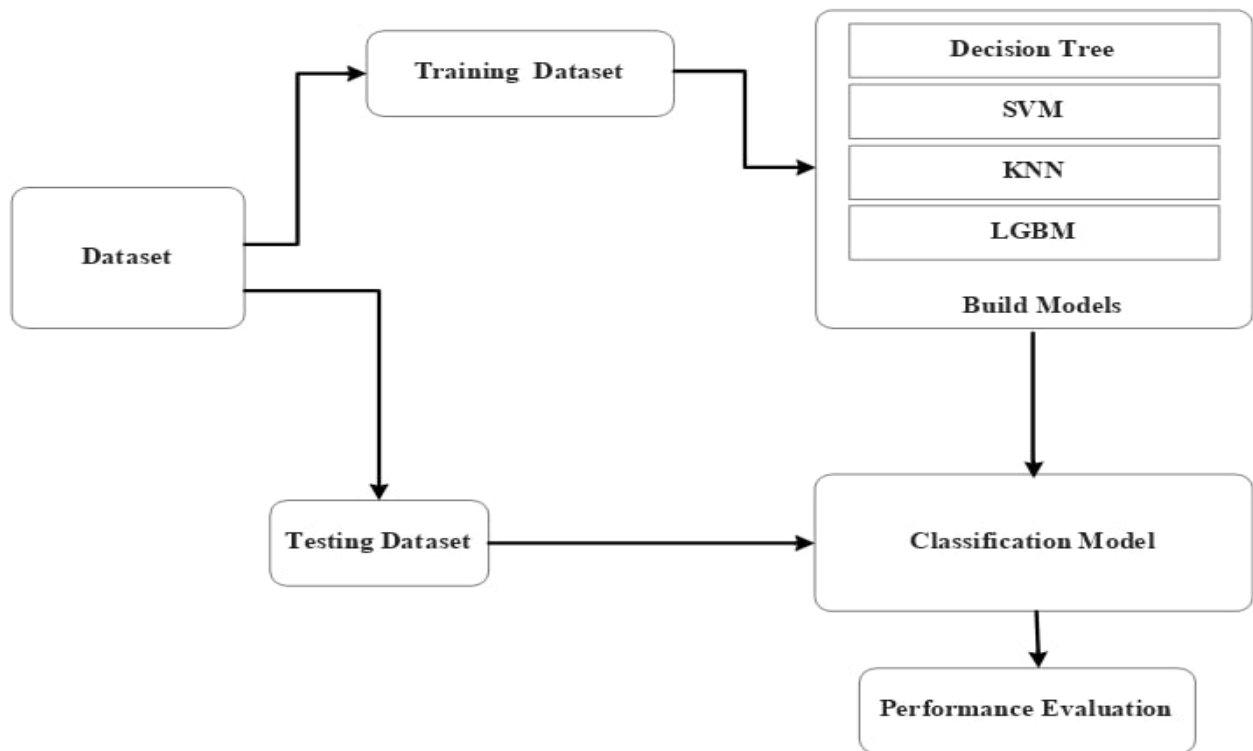


Figure 2: Proposed Models

Step 1: The data set is divided in to testing and training data

Step 2: The training data train models

Step 3: The testing data tests the trained data

Step 3: The loan application is subjected to the trained machine learning model

Step 4: The chance of default is calculated using a machine learning performance

Confusion Matrix

One of the methods to evaluate the performance of a classifier is using confusion matrix. The number of instances predicted correctly or incorrectly by a classification model is summarized in a Confusion matrix. Classifier evaluation measures such as Accuracy, Precision, Recall, and F-measure are included in the confusion matrix. A single prediction in the two-class situation with classes yes and no has four alternative outcomes as shown in table. The true positives (TP) and true negatives (TN) are correct classifications.

A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive [18].

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Table 3: Two dimension confusion matrix

Where:-

- True Positive (TP): It refers to the number of predictions where the classifier correctly predicts the positive class as positive.
- True Negative (TN): It refers to the number of predictions where the classifier correctly predicts the negative class as negative.
- False Positive (FP): It refers to the number of predictions where the classifier incorrectly predicts the negative class as positive.

- False Negative (FN): It refers to the number of predictions where the classifier incorrectly predicts the positive class as negative.

The most frequent metric for evaluating the model's performance is accuracy. The degree of closeness of a data mining classifier's prediction to the actual values, either true or false is characterized as its correctness accuracy [18].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad \text{Equation..... 1}$$

The number of true positives (i.e., items correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class (i.e., the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class) equals the precision for the class [18].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{Equation..... 2}$$

Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been) [18].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{Equation..... 3}$$

Similarly, precision and recall are defused for the negative class. The ratio of the negative over total instance classified as negative can give use precision. In other way the ratio of true negative over total instances of negative class give uses recall. The final metric used for performance evaluation of classifiers on confusion matrix is F-measure. F- Measure is the inverse relationship between precision & recall, and calculated as the harmonic mean of recall and precision. It is the point to conclude that the precision and recall of the model are significantly balanced [18].

$$\text{F - Messure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Pricision} + \text{Recall}} \quad \text{Equation..... 4}$$

3. Results and Discussion

SMOTE (Synthetic Minority Over-sampling Technique) is a technique for learning from data sets that are unbalanced. Based on the SMOTE algorithm, this work proposes a novel way for learning from imbalanced data sets. The application of SMOTE over sampling is covered in this section to demonstrate how this technique is employed and to demonstrate the outcome of improvement [19].

Decision tree

3527 datasets were used to test the decision tree model during this testing. After SMOTE is implemented, this number of data sets is employed, and the result shows that the accuracy of the Decision Tree has improved from 94 percent to 96.79 percent. Precision, recall, and F1 score have all improved as well.

```
DecisionTreeClassifier(max_depth=6, splitter='random')
      precision    recall  f1-score   support

      0         1.00      0.94      0.97         1764
      1         0.94      1.00      0.97         1763

   accuracy                0.97         3527
  macro avg              0.97      0.97      0.97         3527
 weighted avg              0.97      0.97      0.97         3527

Confusion Matrix:
[[1651  113]
 [   0 1763]]
Accuracy of the model on Testing Sample Data: 0.97

Accuracy values for 5-fold Cross Validation:
[0.99619207 0.99885881 0.99923932 0.94824681 0.97870536]

Final Average Accuracy of the model: 0.98
```

Figure 3: Decision tree test result using oversampling (SMOTE)

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Dataset Size} = (1651 + 1763) / 3527 = 0.9679 = \underline{\underline{96.79\%}}$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = 1651 / (1651 + 113) = 1651 / 1764 = 0.9359 = \underline{\underline{93.59\%}}$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) = 1651 / (1651 + 0) = 1651 / 1651 = 1 = \underline{\underline{99.99\%}}$$

$$\text{F-Measure} = 2 * (0.9359 * 0.9999) / (0.9359 + 0.9999) = 0.9679 = \underline{\underline{96.79\%}}$$

KNN

After the dataset was balanced using the SMOTE approach, this model was tested and improved. The model is improved from 90 percent to 94 percent during this testing. As a result, KNN receives a 95 percent model accuracy rating and a 6% misclassification rating.

```
KNeighborsClassifier(n_neighbors=3000)
      precision    recall  f1-score   support

      0       0.99      0.89      0.94       1764
      1       0.90      0.99      0.95       1763

   accuracy                   0.94       3527
  macro avg       0.95      0.94      0.94       3527
 weighted avg       0.95      0.94      0.94       3527

Confusion Matrix:
[[1575  189]
 [   13 1750]]
Accuracy of the model on Testing Sample Data: 0.94

Accuracy values for 5-fold Cross Validation:
[0.92953557 0.95824319 0.93438271 0.95800291 0.96415476]

Final Average Accuracy of the model: 0.95
```

Figure 4: KNN test result using oversampling (SMOTE)

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Dataset Size} = (1575 + 1750) / 3527 = 0.9446 = \underline{\underline{94.46\%}}$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = 1575 / (1575 + 189) = 1575 / 1764 = 0.8928 = \underline{\underline{89.28\%}}$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) = 1575 / (1575 + 13) = 1575 / 1588 = 0.9918 = \underline{\underline{99.18\%}}$$

$$\text{F-Measure} = 2 * (0.8928 * 0.9918) / (0.8928 + 0.9918) = 0.9388 = \underline{\underline{93.88\%}}$$

SVM

After the SMOTE method is applied, the support vector machine (SVM) is also tested. As a result, the testing result has increased from 90 percent to 98 percent. After SMOTE is included in the dataset, SVM model accuracy earns the second highest accuracy throughout this test. As a result, this model only has a 2% misclassification rate.

```

SVC()
      precision    recall  f1-score   support

     0       1.00      0.95      0.97       1764
     1       0.95      1.00      0.97       1763

 accuracy          0.97       3527
  macro avg       0.97      0.97      0.97       3527
 weighted avg     0.97      0.97      0.97       3527

Confusion Matrix:
[[1677   87]
 [    8 1755]]
Accuracy of the model on Testing Sample Data: 0.973

Accuracy values for 5-fold Cross Validation:
[0.97035502 0.98122521 0.97599902 0.97618287 0.98433554]

Final Average Accuracy of the model: 0.98

```

Figure 5: SVM test result using oversampling (SMOTE)

Accuracy = (TP+TN)/Dataset Size= (1677+1755)/3527 =0.9730 = **97.30%**

Precision = TP/ (TP+FP) = 1677/ (1677+87) = 1677/1764=0.9506 = **95.06%**

Recall = TP/ (TP+FN) = 1677/ (1677+8) = 1677/1685 = 0.9952 = **99.52%**

F-Measure = 2*(0.9506 *0.9952)/ (0.9506 +0.9952) = 0.9723= **97.23%**

LightGBM

During the typical dataset testing, the LightGBM model received the highest score. This model has a 95 percent accuracy rating. However, during dataset balancing, this model achieves 99.97% accuracy, and it remains the leading model in this test model. The implementation of the SMOTE model, which will help to over sample the data set, is one of the key reasons for this model improvement. This strategy significantly increased the accuracy, precision, recall, and f1 score of each model.

```

LGBMClassifier(class_weight='balanced', max_depth=-20, n_estimators=4)
      precision    recall  f1-score   support

      0       1.00      1.00      1.00      1764
      1       1.00      1.00      1.00      1763

   accuracy               1.00      3527
  macro avg       1.00      1.00      1.00      3527
 weighted avg       1.00      1.00      1.00      3527

[[1763    1]
 [    0 1763]]
Accuracy of the model on Testing Sample Data: 1.0

Accuracy values for 5-fold Cross Validation:
[1.         1.         1.         1.         0.99961957]

Final Average Accuracy of the model: 1.0

```

Figure 6: LightGBM test result using oversampling (SMOTE)

Accuracy = (TP+TN)/Dataset Size= (1763+1763)/3527 =0.9997 = **99.97%**

Precision = TP/ (TP+FP) = 1763/ (1763+1) = 1763/1764=0.9994 = **99.94%**

Recall = TP/ (TP+FN) = 1763/ (1763+0) = 1763/1763= 0.9997= **99.98%**

F-Measure = $2 * (0.9994 * 0.1) / (0.9994 + 0.1) = 0.9998 = \mathbf{\underline{99.98\%}}$

Summary of testing result

The table below summarizes the model testing results after the data set has been balanced using the oversampling SMOTE approach. The results of each model are greatly boosted and improved after SMOTE is implemented in this dataset, indicating that data balancing has an impact until correct dataset balancing is implemented. So, after implementing oversampling SMOTE methode, the accuracy of the LightGBM model has increased to 99.97%, SVM has improved to 97.30%, Decision tree has improved to 96.79%, and KNN has improved to 94.46 percent. The outcomes of tested models utilizing the newly balanced dataset that have been changed using the SMOTE technique are described in the table below.

Parameters	Predictions on New Data			
	LGBM	DTree	SVM	KNN
Total number of instances (test data set)	3527	3527	3527	3527
Correctly classified instance	3526	3414	3432	3325
Incorrectly classified instance	1	113	95	202
Accuracy	99.97	96.79	97.30	94.46
Precision	99.94	93.59	95.06	89.28
Recall	99.98	99.99	99.52	99.18
F-Measure	99.98	94	97.23%	93.88%

Table 4: Summery of tested model using balanced dataset

The graph below illustrates the comparison of the four models after over sampling of SMOTE method is implemented; as indicated in the table above; both parameters (accuracy, precision, recall, and f-measures) are defined in numeric form. The graph below depicts the information in the form of a graph for easier understanding.

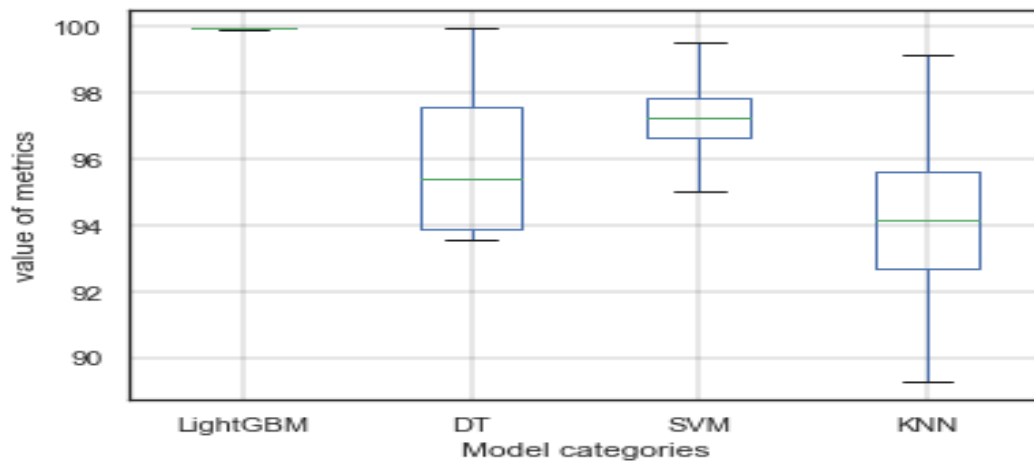


Figure 7: Graphical comparisons of models after SMOTE is implemented

Summary of Models Comparison

This study aims to minimize credit risks at Ethiopian Micro finance Institutions by identifying potential borrowers using machine learning models. Various models were tested on both balanced and unbalanced datasets. The results show that all models performed better with balanced data, with LightGBM achieving the highest improvement from 96% to 99.97%, followed by SVM, Decision Tree, and KNN. LightGBM emerged as the best-performing model.

Parameters	Models testing using normal dataset				Predictions on New Data			
	LGBM	DTree	SVM	KNN	LGBM	SVM	DTree	KNN
Total number of instances (test data set)	2630	2630	2630	2630	3527	3527	3527	3527
Correctly classified instance	2509	2471	2430	2379	3526	3432	3414	3325
Incorrectly classified instance	121	159	200	251	1	95	113	202
Accuracy	95	94	90	92.3	99.97	97.30	96.79	94.46
Precision	99	94	93	87	99.94	95.06	93.59	89.28
Recall	86	98	83	84	99.98	99.52	99.99	99.18
F-Measure	93	90	88	85	99.98	97.23	94	93.88

Table 5: Summery of model comparisons with unbalanced and balanced dataset

Discussion

The discussion centers on the processes and performance evaluations of four predictive models—LightGBM, SVM, Decision Tree, and KNN used to assess credit risk in Ethiopian Micro finance Institutions. The models were built and tested using the SMOTE technique to balance the datasets. The LightGBM model outperformed others, accurately classifying 99.97% of the data with a false positive rate of 0.02%. SVM followed with 97.30% accuracy and a false positive rate of 0.047%, while Decision Tree and KNN achieved 96.79% and 94.46% accuracy, respectively, with corresponding false positive rates of 0.06% and 0.097%. The study emphasizes the importance of minimizing false positives, which represent credits incorrectly classified as non-defaults, as this could result in financial losses for creditors. False negatives, which involve classifying defaults as non-defaults, were minimal across the models, with LightGBM and Decision Tree achieving 0% false negatives, SVM 0.007%, and KNN 0.004%. Overall, LightGBM achieved the highest accuracy, precision, recall, and F1 score, followed by SVM and Decision Tree, with KNN trailing. The study's

findings are more accurate than those in previous research due to the larger dataset of 13,148 records and 12 attributes, as well as the use of the SMOTE oversampling method. Misclassification rates, which represent the percentage of incorrect predictions, were also discussed, with recommendations for improving accuracy and reducing errors through balanced sampling and careful tuning of model parameters.

Summery

This study employed four machine learning models—LightGBM, Decision Tree, KNN, and SVM—to predict credit risk by classifying potential borrowers. The study aimed to answer three key research questions.

1. **Major Influential Factors:** The study identified key factors influencing credit risk prediction, including 'Saving Balance', 'Loan Amount', 'Paid Principal', 'Marital Status', 'Sex', 'Other Source of Income', 'Educational Level', 'Collateral', 'Business Sector', 'Loan Term', 'Credit History', and 'Loan Status'. These features were selected using the attribute correlation method.
2. **Model Accuracy:** Among the models, LightGBM emerged as the most accurate for forecasting credit risk, outperforming the others due to its ability to handle large datasets efficiently with high accuracy. KNN was the least accurate. LightGBM's histogram-based technique enables faster training speeds and requires less memory, making it ideal for large and complex datasets.
3. **Model Development and Evaluation:** The study involved cleaning the data, choosing the appropriate predictive modeling approach, preprocessing the data, training the model with 80% of the dataset, testing the model's performance, validating its accuracy on unseen data, and using the model for prediction once satisfied with its performance.

The study's comprehensive evaluation of these models helps answer the research questions and supports the development of effective credit risk prediction tools.

Conclusion

This study evaluated the use of machine learning models to assess individuals' credit risk in a microfinance context, particularly at Ethiopian Micro Finance Institutions. Traditional models like linear regression, while still in use, have been outperformed by machine learning models, which provide more accurate predictions. The study focused on classifying loan applicants as either Active or Default using real data, emphasizing the need for managers to rely on machine learning to simplify and filter vast amounts of data.

The study followed a standard process of data preparation, including selection, preprocessing, cleanup, data mining, and result validation. Using a dataset of 13,148 records, the data was divided into 80% for training and 20% for testing, employing a 5-fold cross-validation approach. Four machine learning algorithms—LightGBM, SVM, Decision Tree, and KNN—were used to build prediction models, with tools like Python and Jupyter for simulation.

The LightGBM model achieved the highest accuracy at 99.97%, followed by SVM at 97.30%, Decision Tree at 96.79%, and KNN at 94.46%. From 21 initial attributes, 12 were selected as key predictors, including Savings Balance, Paid Principal, Age, Marital Status, and Loan Status.

The study concludes that these machine learning models can help Ethiopian Micro Finance Institutions and optimize their loan approval process, identifying high-risk applicants more effectively. However, the study acknowledges limitations, such as the need for larger datasets and consideration of temporal factors in future research to enhance prediction accuracy.

Reference

- [1]. Krivoruchko, S.V., Abramov, M.A., Mamut, M.V., Webspinner, O.S., Shaker, I.E. (2012), the risks of microfinance and regulation. In: Krivoruchko, S.V., Abramov, M.A., Mamut, M.V., Webspinner, O.S., Shaker, I.E., editors. Microfinance in Russia. Moscow: Center for Research Payment and Settlement Systems.
- [2]. A.Gebrehiwot, "Micro-finance institutions in Ethiopia: issues of portfolio risk, institutional arrangements and governance," Ethiopian Journal of Economics, vol.
- [3]. E. Salame, "Applying data mining techniques to evaluate applications for agricultural loans," 2011.
- [4]. S. Sirikulvadhana, "Data mining as a financial auditing tool," Citeseer, 2002.
- [5]. J. Wu, S. Vadera, K. Dayson, D. Burridge, and I. Clough, "A comparison of data mining methods in microfinance," in Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on, 2010, pp. 499-502.
- [6]. A. S. Koyuncugil and N. Ozgulbas, "A Data Mining Model for Detecting Financial and Operational Risk Indicators of SMEs," World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, vol. 2, pp. 1082-1085, 2008.
- [7]. P Salman Raju, Dr V Rama Bai, and G Krishna Chaitanya, (2014) Data mining: Techniques for. Enhancing Customer Relationship Management in Banking and microfinance institutions 2004
- [8]. A. J. Deneke, "The application of data mining to support customer relationship management at Ethiopian airlines," 2003.
- [9]. Sara Worku, "Application of Data Mining Technology for Credit Risk Assessment in Addis Credit and Saving Institution, "Master's Thesis, Addis Ababa University, June 2016.
- [10]. Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance 34: 2767–87.
- [11]. Yap, Bee Wah, Seng Huat Ong, and Nor Huselina Mohamed Husain. 2011. Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications 38: 13274–83.
- [12]. Zhao, Zongyuan, Shuxiang Xu, Byeong Ho Kang, Mir Md Jahangir Kabir, Yunling Liu, and Rainer Wasinger. 2015. Investigation and improvement of multi-layer perceptron neural networks for credit scoring. Expert Systems with Applications 42: 3508–16.

- [13]. Alex Addae-Korankye,” Causes and Control of Loan Default/Delinquency in Microfinance Institutions in Ghana” American International Journal of Contemporary Research Vol. 4, No. 12; December 2014. 66
- [14]. Dahee Choi, Kyungho Lee an Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation” Hindawi Security and Communication Networks Volume 2018, Article ID 5483472,25 September 2018.
- [15]. Efficient feature selection filters for high-dimensional data by Artur J. Ferreira , Mário
- [16]. Sara Worku, “Application of Data Mining Technology for Credit Risk Assessment in Addis Credit and Saving Institution, “Master’s Thesis, Addis Ababa University, June 2016.
- [17]. Maimon, O.; Rokach, L. Data Mining and Knowledge Discovery Handbook; Springer: New York, 2005. [Crossref], Google Scholar
- [18]. D. L. Gupta, A. K. Malviya, Satyendra Singh,” Performance Analysis of Classification Tree Learning Algorithms,” International Journal of Computer Applications (0975 – 8887) Volume 55– No.6, October 2012.
- [19]. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)